2016

Learning hierarchical 3D kernel descriptors for RGB-D action recognition

Kong, Yu

Computer Vision and Image Understanding 144 (2016) 14–23

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Learning hierarchical 3D kernel descriptors for RGB-D action recognition

Yu Kong [a,*], Behnam Satarboroujeni [a], Yun Fu [a,b]

[a] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA
[b] College of Computer and Information Science, Northeastern University, Boston, MA, USA

ARTICLE INFO

Article history:
Received 22 December 2014
Accepted 1 October 2015

Keywords:
RGB-D action
Action recognition
Kernel descriptor

ABSTRACT

Human action recognition is an important and challenging task due to intra-class variation and complexity of actions which is caused by diverse style and duration in performed action. Previous works mostly concentrate on either depth or RGB data to build an understanding about the shape and movement cues in videos but fail to simultaneously utilize rich information in both channels. In this paper we study the problem of RGB-D action recognition from both RGB and depth sequences using kernel descriptors. Kernel descriptors provide an unified and elegant framework to turn pixel-level attributes into descriptive information about the performed actions in video. We show how using simple kernel descriptors over pixel attributes in video sequences achieves a great success compared to the state-of-the-art complex methods. Following the success of kernel descriptors (Bo, et al., 2010) on object recognition task, we put forward the claim that using 3D kernel descriptors could be an effective way to project the low-level features on 3D patches into a powerful structure which can effectively describe the scene. We build our system upon the 3D Gradient kernel descriptor and construct a hierarchical framework by employing efficient match kernel (EMK) (Bo, and Sminchisescu, 2009) and hierarchical kernel descriptors (HKD) as higher levels to abstract the mid-level features for classification. Through extensive experiments we demonstrate the proposed approach achieves superior performance on four standard RGB-D sequences benchmarks.

1. Introduction

Despite the many research efforts in recognizing human actions and many encouraging advances in computer vision field, still accurate action recognition is a challenging task due to large intra-class variation and background noise. Introducing low-cost devices such as Kinect sensors has triggered many research activities for achieving concise descriptions in recognition task due to their availability of depth sequences alongside the RGB data. Insensitivity of depth images to different lighting situations and illuminations is an effective advantage compared to color images. Moreover, depth sequences provide additional shape and movement information due to providing accurate distance information for each pixel in image.

However, depth sequences [5,6,27] are noisy with undefined depth data and incorrect joint data. Existing work [4,6,27] directly builds on low-level noisy features, which are hardly being linearly separated, and thus this would hurt the performance due to the noisy raw features. In addition, the correlations between human body parts are highly nonlinear. It is difficult to model their joint distribution accurately by extracting features from each of them and concatenating the two features.

It would be better to utilize other robust source of information to alleviate the problem of noisy data. In using depth sequences, recent works such as [4] consider the human motion as a posture of the body segments and employ the skeleton tracker to construct a discriminative representation from depth sequences. However, skeleton data are also noisy, and the correlations between skeleton data and depth data are highly nonlinear and difficult to learn. In addition, as presented in [27], utilizing low-level attributes in depth images outperforms recent high-level representations with improvement on capturing joint shape-motion cues. This idea leads us towards employing low-level attributes in depth images in a more elaborate way to capture accurate information in describing action scene.

In this paper, we propose a hierarchical kernel based method to learn the non-linear correlations between RGB and depth action data for action recognition. The aforementioned problems are overcome by a novel hierarchical kernel framework motivated by the recent success of kernel descriptors for object recognition task [2]. We propose a 3D gradient kernel descriptor which is a low-level depth sequence descriptor with ability of capturing detailed information by computing pixel-level 3D gradient. The framework of our approach is illustrated in Fig. 1. Our 3D gradient kernel is essentially producing normal vectors on the surface of 3D geometric shape of scene using

* Corresponding author.
E-mail address: yukong@ece.neu.edu (Y. Kong).

http://dx.doi.org/10.1016/j.cviu.2015.10.001
1077-3142/

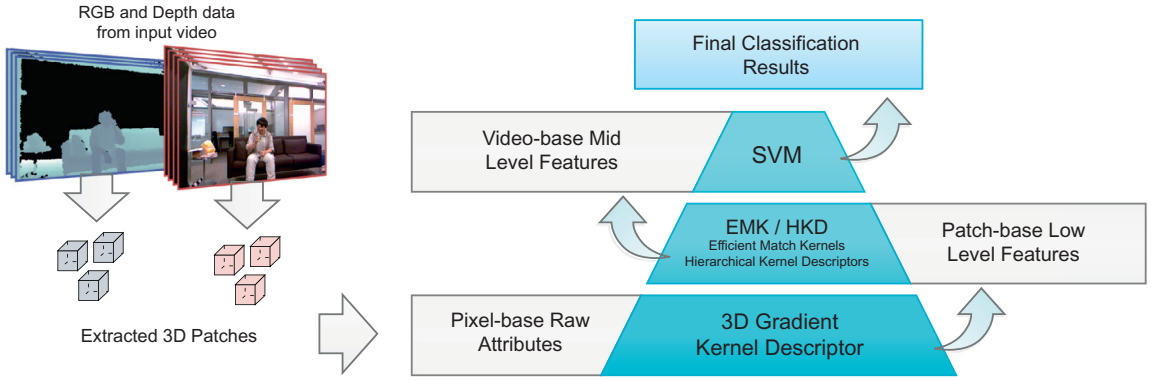

Fig. 1. Hierarchical framework for learning mid-level features from depth and RGB sequences. Our 3D kernel descriptor extracts low-level feature from 3D patches, then we use EMK or HKD to abstract mid-level feature for feeding the classifier.

the gradient in depth images. Moreover, the gradient information is computed along the temporal dimension as well as the spatial dimensions to describe the change in shape of the 3D surface in time. As it was shown in [7], using the normal vectors in depth images provides a rich description of the scene.

At higher level, we use two different methods to summarize the mid-level features for classification. The efficient match kernels (EMK) is the first method which allows us to learn nonlinear correlations between body parts. The learned high-order correlations accurately measure the similarities between two RGB-D action videos, and provide rich mid-level information to bridge the semantic gap for classification. We also apply hierarchical kernel descriptors (HKD) as higher level of our framework to aggregate patch-level features into one representative feature vector for each video. These hierarchical kernel descriptors are based on the first layer of our framework where instead of working on pixel-level attributes, they get the patch-level features as input and generate a feature vector for each patch block. Finally, the classification is performed by using linear SVM.

This work is an extension of our previous paper [1]. The extensions are: (a) Type of data, here we use both RGB and depth data while in [1] we just focus on depth data. (b) Method used in middle layer of framework. We just use EMK in [1] while here we introduce HKD and show how it improves our results. (c) Extensive experimental result on more datasets.

Our work differs from existing normal-based methods [7,27,30]. One major difference is that our method utilize rich information in both color and depth sequences, while [7,27,30] only uses depth information. In addition, we compute nonlinear efficient match kernels for a RGB-D video, while [7,27,30] aggregates local features to build the representative feature vector for each video. Our method is particularly designed for RGB-D actions, while [2,3] were designed for object recognition. We compute 3D gradient in the low-level feature extraction to capture temporal information, while [2,3] uses 2D gradient. We achieve state-of-the-art results on three public RGB-D action datasets, and comparable results on the fourth dataset, while [2,3] was not tested on RGB-D action recognition datasets. We carefully adapted the parameters in the 3D gradient, EMK and HKD, and show the optimal parameters for RGB-D action recognition, as the proposed method is sensitive to those parameters.

We show how our framework is applied to RGB-D sequences for achieving discriminative information and surpassing sophisticated learning approaches based on high-level features. Our method is extensively evaluated on four RGB-D datasets, MSR Action 3D [8], MSR Gesture 3D [6], MSR Action Pairs [27], and MSR DailyActivity 3D datasets [4], and show superior performance over state-of-the-art approaches.

2. Related work

Action recognition has been widely explored throughout the computer vision community. Early work utilized preexisting color sequence methods as a basis to exploit depth sequences. These early attempts can be split into two categories, low-level feature-based methods and mid-level knowledge-based methods. Low-level feature-based methods [9–13,42] adopt spatio-temporal interest points, body shape feature [14], structure information [15], key poses [16], etc., to represent human actions. These methods use hand-crafted features to learn actions. Recent work show that action features can also be learned using deep learning techniques [17].

Regardless of the progress mentioned above, we still have not found a promising representation of actions that connects low-level features and high level semantics. In response to this, mid-level features, such as attributes [18], body parts [19], semantic descriptions [20], and context [21], are summarized from low-level features, and then used for action classification. These mid-level features can be either discovered from the training data or generated by a human expert.

Compared with these action recognition approaches that designed for RGB videos, we propose a novel framework that elegantly fuses RGB and depth modality data for RGB-D action recognition. Due to the unavailability of depth data, previous approaches for RGB video classification do not have the ability to use 3D structural information of the entire scene, and thus confuse actions. By comparison, we use depth data that inherit 3D structural information. This crucial information can be effectively used to simplify intra-class motion variations and remove cluttered background noise. Furthermore, our approach is an unified feature extraction algorithm that is suitable for heterogeneous RGB and depth data. It is modality-free and can be applied to both RGB and depth videos.

Recently, due to introducing the cost-effective Kinect sensor to the market, researchers have devoted great efforts to recognize actions with RGB-D data [22–27]. Compared with conventional RGB data, the extra depth data allow us to capture 3D structural information, which reduces background information and simplifies intra-class motion variations. Some works [28] also attempted to reconstruct the depth information in datasets containing only RGB data. Some of early research works on RGB-D data were based on treating the new depth information as just another type of 2D

information. Employing old descriptors on depth data and using the extracted features alongside the RGB features was too explored [26]. Although some methods [29] extract additional information specifically for depth data. Among these methods, as demonstrated in [7,30] surface normals provide rich information about shape and structure in depth sequences. HON4D [27] followed the same method by extending the surface normals to a 4D space, and then quantized them to achieve discriminative information about the scene.

Our work also uses the concept of surface normal. We employ it alongside with kernel descriptors to eliminate the loss of information during quantization. Kernel descriptors are easy to design and outperform methods that quantize continuous data by gathering additional aspects of the input data for accurate presentation in learning phase.

Our approach differs from local feature-based approaches [25,44] and skeleton-based approaches [43,45]. We learn hierarchical kernel descriptors, which are essentially nonlinear feature mappings. Our approach can better describe spatiotemporal structures, and it was demonstrated in our experiments. The feature learning approach proposed in our paper captures holistic motion of the entire scene including both human and background. In contrast, approaches in [25,44] extract local spatiotemporal features. Our approach is capable of fusing RGB and depth data for classification while methods in [25,44] can only be applied to depth data. Compared with skeleton data-based approaches [43,45], we propose an elegant framework that is capable of fusing RGB and depth data while it is not clear how the evolutionary algorithm [43] can fuse various modality data. We achieve comparable results with [44,45] and outperform [43] on MSR-Action3D dataset as RGB videos are not provided in this dataset. We believe our results can be further boosted if RGB videos are available.

3. Our method

Given a sequence of depth images, we design a match kernel for gathering pixel-level gradient information, which is equivalent to surface normals [7] in 3D geometric view. Following [2], we learn compact basis vectors via KPCA [31], then we build our kernel descriptor with projecting infinite-dimension feature vectors which are generated by our 3D gradient match kernel, into a finite set of basis vectors.

3.1. Kernel descriptors

Kernel descriptors have been demonstrated to be very effective in capturing pixel-level features compared to other methods such as SIFT [32] and HOG [33], as they are able to gather more descriptive information lying in high dimensional space. In this paper we present a 3-dimensional kernel descriptor over normal vectors on the 3D depth surface with expanding the current kernel descriptors for object recognition task in 2D images.

Based on the description in [2], 3D gradient match kernel

$$K_{3D}(P,Q)=\sum_{z\in P}\sum_{z'\in Q}\widetilde{m}_z\widetilde{m}_{z'}k_o(\theta_z,\theta_{z'})k_p(z,z') \tag{1}$$

is the kernel representation of orientation histograms over video patches, where P and Q are the 3D patches from different videos and the overall output is a measurement of similarities between them. Let m_z and θ_z be the 3D magnitude and orientation of gradient at the pixel z in a patch, then $\widetilde{m}_z=m_z/\sqrt{\sum_{z\in P}m(z)^2+\epsilon}$ is the normalized magnitude of gradient which is used as a weight factor. Gaussian orientation kernel $k_o(\theta,\theta')=\exp(-\gamma_o\|\theta_z-\theta_{z'}\|^2)=\phi_o(\theta_z)^\top\phi_o(\theta_{z'})$ is for measuring similarities between orientations of gradient in corresponding pixels, and similarly $k_p(z,z')=\exp(-\gamma_p\|z-z'\|^2)=\phi_p(z)^\top\phi_p(z')$ is the Gaussian kernel for measuring how close each pair of pixels lay in 3D spatio-temporal dimension of each patch.

In case of depth sequences, we have a richer description in each frame compared to RGB images. Because alongside of spatial location of each pixel, we also have a third dimension which is the distance of each pixel from camera, shown as intensity values. This fact helps us to capture the shape of 3D geometric surface by considering the orientation and magnitude of 2D gradient in each pixel of a depth map. With considering 3D gradient alongside of spatio-temporal dimension we are capturing the change of shape of this surface over time, which is an essential factor for having a rich discriminative representation of each 3D patch.

3.2. Feature extraction

The 3D gradient match kernel uses two patches and gives a measurement of similarities between them. However, the goal of kernel descriptors is to produce an independent discrimination over each individual patch. With expanding the orientation and position kernels in Eq. (1), the extracted features over each patch is represented as

$$F_{3D}(P)=\sum_{z\in P}\widetilde{m}(z)\phi_o(\theta(z))\otimes\phi_p(z), \tag{2}$$

where $\otimes$ is the Kronecker product. Instead of measuring the similarities between two patches, $F_{3D}(P)$ is only depended on one patch and is used as the feature vector. Extracting the output of Eq. (1) over patches P and Q is done by $K_{3D}(P,Q)=F_{3D}(P)^\top F_{3D}(Q)$. However, our goal is to extract individual features based on $F_{3D}(P)$. Presence of Gaussian kernel in formulation of $F_{3D}(P)$ introduces infinite dimensionality. Therefore, computing $F_{3D}(P)$ is infeasible.

For reducing the dimensionality we project the $F_{3D}(P)$ to a finite set of basis vectors. In choosing an efficient set of basis vectors following the presented method in [3], we use a fine grid over the support region for approximation. Let $\{\phi_p(x_i)\}_{i=1}^b$ be the set of basis vectors where b is the number of basis vectors and x_i is the sample normalized vector used for approximation of Gaussian kernel $k_p(z,z')$ over position of pixels in 3D space. The mechanism of projecting infinite-dimension vector $\phi_p(z)$ to the low-dimensional basis vector set $\{\phi_p(x_i)\}_{i=1}^b$ is equivalent to using the following finite dimensional kernel:

$$\widetilde{k}_p(z,z')=k_p(z,X)^\top[\boldsymbol{K_p}^{-1}]_{ij}k_p(z',X)=[\boldsymbol{G}k_p(z,X)]^\top[\boldsymbol{G}k_p(z',X)], \tag{3}$$

where $k_p(z,X)=\{[k_p(z,x_1),\ldots,k_p(z,x_b)]\}^\top$ is a vector with size equal to number of basis vectors, $K_{pij}=k_p(x_i,x_j)$ is a square matrix with b dimensions, and $K_p^{-1}=\boldsymbol{G}^\top\boldsymbol{G}$. By following Eq. (3) and writing the same for orientation kernel with b_o basis vectors, we have finite-dimension feature vector $\widetilde{F}_{3D}(P)=\sum_{z\in P}m(z)\widetilde{\phi}_o(\theta(z))\otimes\widetilde{\phi}_p(z)$, where $\widetilde{\phi}_o(\theta(z))=\boldsymbol{G}k_o(\theta_z,X)$ with only b_o dimension and $\widetilde{\phi}_p(z)=\boldsymbol{G}k_p(z,X)$ with b_p dimension.

3.3. Dimensionality reduction

Presence of Kronecker product in producing the feature vector alongside with using grid approximation on 3D space make $\widetilde{F}_{3D}(P)$ to have a high dimensionality. In particular we quantize the position kernel k_p with basis vectors on a $4\times4\times4$ grid, and gradient orientation kernel k_o with basis vectors on a $6\times6\times6$ grid in all experiments. Therefore, the final dimensionality of $\widetilde{F}_{3D}(P)$ is $64\times216=13,824$. Although we project $F_{3D}(P)$ to finite dimension now, the dimensionality is still too high for empirical use.

For dealing with aforementioned problem and handling the computation cost, we use the formulation in [2] and try to project our feature vector to a set of joint basis vectors $\{\phi_o(x_1)\otimes\phi_p(y_1),\ldots,\phi_o(x_{b_o})\otimes\phi_p(y_{b_p})\}$, where $\{\phi_o(x_i)\}_{i=1}^{b_o}$ and $\{\phi_p(y_i)\}_{i=1}^{b_p}$ are the set of basis vectors of orientation and position kernels approximation accordingly.

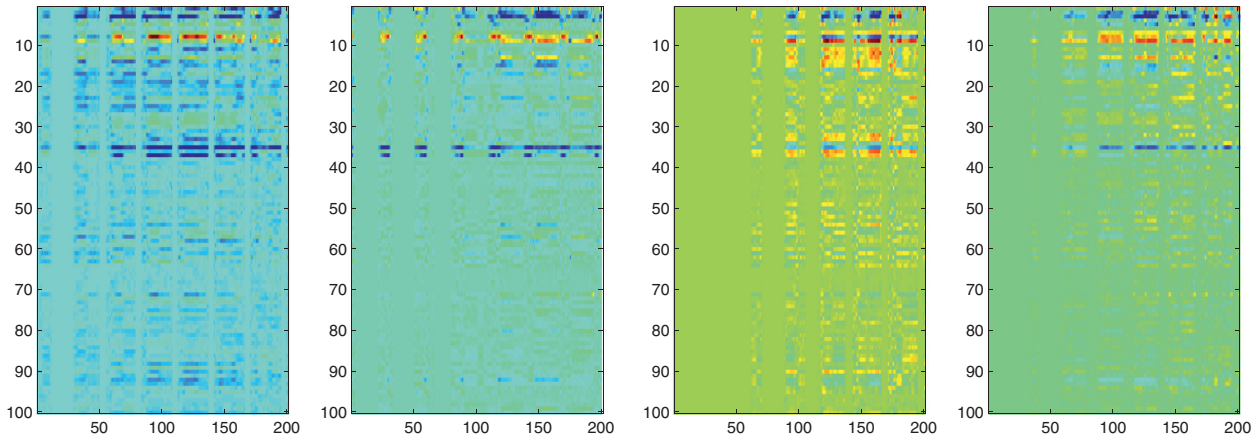

Fig. 2. Instances of output feature vectors on 200 patches from four different action classes. Extracted feature vector for each patch is shown in columns.

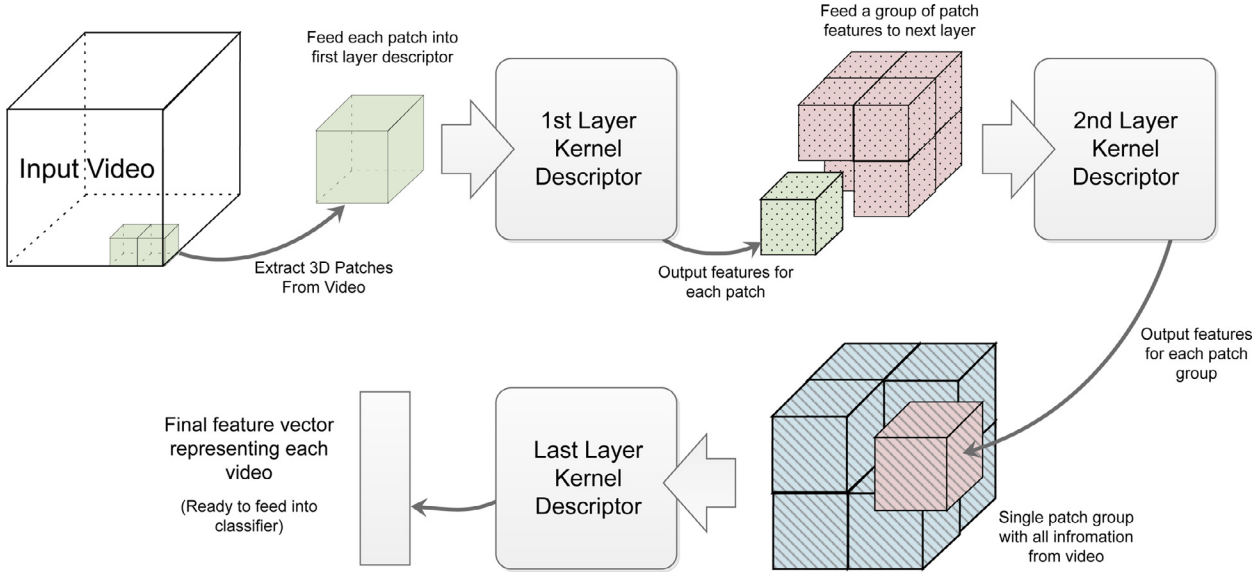

Fig. 3. Framework of hierarchical kernel descriptors (HKD) in overall structure. In each level we use output of previous layer and treat is like pixel-level attribute in first layer. A group of patch features are passed to next layer as a single patch.

3.4. Efficient match kernel (EMK)

Our 3D Gradient kernel descriptor produces low-level discriminative features over each 3D patch in video. Fig. 2 shows the representation of feature vectors from four different action classes. Variability of feature vector patterns expresses the ability of our method to discriminate between various categories. In construction of our hierarchical structure, we employ EMK as the second layer over the output of our 3D Gradient kernel descriptor to abstract mid-level features for classification.

Similar to the concept of kernel descriptors, EMK [3] is the kernel representation of well-known bag-of-words method which has been shown to produce more accurate quantization and a better performance as a result.

One way to use match kernels as an alternative to bag-of-words (BOW) method is by adding local kernels over all combination of local features from two different samples (Sum kernels[34]). This method and lots of other approaches to employ kernels in this manner suffer from space and time complexity as they need to evaluate the full kernel matrix. EMK does not require the explicit computation of a full kernel matrix which makes its computation complexity linear in both time and space. In EMK local features are mapped to a low dimensional feature space and set-level features can be constructed by averaging the resulting feature vectors.

In BOW, each local feature is quantized into a D-dimensional binary indicator vector $\mu(x) = [\mu_1(x), \ldots, \mu_D(x)]^\top$. $\mu_i(x)$ is 1 if $x \in R(v_i)$ and 0 otherwise, where $R(v_i) = \{x : \|x - v_i\| \leq \|x - v\|, v \in V\}$. The feature vectors for one image will be a normalized histogram $\mu(X) = \frac{1}{|X|}\sum_{x\in X}\mu(x)$ where $|.|$ is the cardinality of a set. BOW features can be used together with either a linear or a kernel classifier to perform the classification task. The resulting kernel function using a kernel classifier is:

$$K_B(X,Y) = \bar{\mu}(X)^\top \bar{\mu}(Y) = \frac{1}{|X||Y|}\sum_{x\in X}\sum_{y\in Y}\mu(x)^\top\mu(y)$$

$$= \frac{1}{|X||Y|}\sum_{x\in X}\sum_{y\in Y}k(x,y)$$

where $k(x, y)$ is a positive definite kernel, measuring the similarity between two local features x and y.

3.5. Hierarchical Kernel descriptors (HKD)

We also expand our experiments by using another method as higher levels in our framework using same kernel descriptor as the first layer. In contrast with the first layer, here we try to use previous layer data as input instead of pixel-level gradient. Fig. 3 shows the framework where after each level we link a group of features as a

single patch and feed it to the next layer until we have a final single feature vector to represent the whole video. Our formulation for HKD is similar to our 3D gradient kernel descriptor:

$$K_{\mathrm{HKD}}(P_G, Q_G) = \sum_{Z\in P_G}\sum_{Z'\in Q_G} \tilde{w}_Z \tilde{w}_{Z'} k_F(F_Z, F_{Z'}) k_p(Z, Z') \quad (4)$$

where P_G and Q_G are the 3D patch groups from different videos which instead of pixel attributes now have output features from preceding layer. Like before we have two Gaussian kernels $k_F(F_Z, F_{Z'}) = \exp(-\gamma_o \|F_Z - F_{Z'}\|^2) = \phi_F(F_z)^\top \phi_F(F_{z'})$ and $k_p(Z, Z') = \exp(-\gamma_p \|Z - Z'\|^2) = \phi_p(Z)^\top \phi_p(Z')$. Here $k_F(F_Z, F_{Z'})$ the Gaussian kernel for computing the similarities between the extracted features from previous layer and $k_p(Z, Z')$ is the Gaussian kernel for measuring how close each pair of patches are in 3D spatio-temporal dimension. We used magnitude of gradient as the weight factor in formulation of first layer descriptor, and similarly here we use average weight of preceding layer components as our weight factor. Therefore in each layer we use the average gradient magnitude of pixel which have contributed to feature extraction in all previous layers. We have $\tilde{w}_Z = w_Z / \sqrt{\sum_{Z\in P_G} w(z)^2 + \epsilon}$ as the normalized weight factor of patch group P_G.

HKD is capable of capturing structure information in the spatio-temporal domain. It performs well when structure information is critical to the recognition performance. We experimentally compare their performance in Section 4 and show HKD outperforms EMK in most of the cases.

4. Experiments

We test our approach on four standard RGB-D activity datasets including MSR Action 3D [5] dataset, MSR Action Pairs [27] dataset, MSR Gesture 3D [6] dataset, and MSR Daily Activity [4] dataset.

4.1. Datasets

MSR Action 3D [5] dataset is an action dataset of depth sequences captured by a depth camera. There are 10 subjects each performing an action for two or three times. All together this dataset has 567 depth sequences with resolution of 320 × 240 containing 20 action categories: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, sideboxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up,* and *throw*. We employ our method on this dataset with standard experiment setup [4], where instances from half of the subjects are used as train data, and the rest of them are used as test data. In particular subjects 1,3,5,7, and 9 are used for training data and the rest are used for test data.

MSR Action Pairs [27] dataset is an action dataset of RGB and depth sequences. Availability of RGB data in this dataset makes it suitable for conducting experiments on both color map and depth sequences. This dataset consists of activities with similar motion and shape cues which makes it challenging for methods such as [4] which are based on motion cues. For example pair of actions *"pick up a box"* and *"put down a box"* have different correlations but are similar in terms of motion. Using temporal relation of frames in these sequences is the key for accurate classification.

There are 12 different actions (6 pairs) in this dataset, consisting of: *"pick up a box/put down a box","lift a box/place a box", "push a chair/pull a chair", "wear a hat/take off a hat", "put on a backpack/take off a backpack",* and *"stick a poster/remove a poster"*. Each action is performed three times by ten subjects which results in 360 instances. The first five subjects are used for training and the rest of them are used for testing.

MSR Gesture 3D [6] dataset is a hand gesture dataset of depth sequences and contains a group of American Sign Language (ASL) gestures. There are 12 gestures in this dataset which represent: *bathroom, blue, finish, green, hungry, milk, past, pig, store, where, j,* and *z*. Each gesture is performed two or three times by 10 different subjects with the hand portion is captured as the final instance in dataset. The dataset contains 336 depth sequences. For recognition in this dataset both shape and movement of hands are important. Self-occlusion is one of the factors which make this dataset a challenging benchmark for action recognition. We test our method on this dataset using cross-subject test where in each test we use the data gathered from one person for testing and the other 9 persons for training.

MSR Daily Activity [4] dataset is also one the datasets with both RGB and depth data. There are 320 instances in 16 types of activity: *drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up,* and *sit down*. This dataset is designed to represent people's daily activities in the living room. There are 10 subjects performing each action twice in two different poses: "sitting on a sofa" and "standing". This makes the dataset very challenging as it presents a large intra-class variation.

In all our experiments with this dataset, we use one person out method. Other publication has not reported the exact split of data for test and training, therefore it is not possible for us to conduct a fair comparison between our results and other methods.

4.2. Experimental settings

Presenting undefined depth points in depth images as black (zero intensity in gray-scale representation) dots make popular interest point detectors such as STIP [35] to perform poorly in detecting discriminative patches in depth sequences. For gathering maximum amount of information and dealing with aforementioned problem, we employ dense sampling over 3D patches throughout the whole video. To handle the computational cost of dense sampling, we resize instances to be no larger than 150 × 150 in spatial dimensions with preserved ratio. We exploit either the efficient match kernels EMK [3] or hierarchical kernel descriptors (HKD) for producing the video-level features in next level of our hierarchical structure. Finally, we use the linear SVM over video-level features for classification task.

Choosing dataset dependent hyperparameters and running empirical tests to get the best parameters can be an effective way for boosting the performance. However we set some of the parameters fixed during all experiments and try to run practical experiments with parameters of EMK where ever we use it as second level of our framework. Kernel parameters in orientation and position kernel are $\gamma_o = 5$ and $\gamma_p = 3$. The ϵ value in computing the normalized gradient magnitude is set to 0.8. We run experiments with different values for number of basis vectors in CKSVD and its kernel value. Fig. 4 shows the overall accuracies on all datasets with changing these values. In using dense sampling, the overall performance is expected to be better with using multiple patch sizes. However, we empirically choose patches with size 16 × 16 × 16 with 50% overlap with neighbor patches in sampling for all experiments.

Using HKD as higher level of our framework requires some preprocessing on input data. In order to achieve similar grouping of patch features in each level we need to have all of our input videos with the same size in both spatial and temporal manner. In all experiments for HKD we resize video dimensions to some coefficient of 8 (sampling size) to have a neat patch grouping convention among input videos.

We also employ bag of words for encoding features to show the difference in using EMK, HKD compare to classic BOW method.

4.3. Results on MSR action 3D dataset

We employ our method on this dataset with standard experiment setup [4], where instances from half of the subjects are used as train data, and the rest of them are used as test data. In particular subjects 1,3,5,7, and 9 are used for training data and the rest are used for

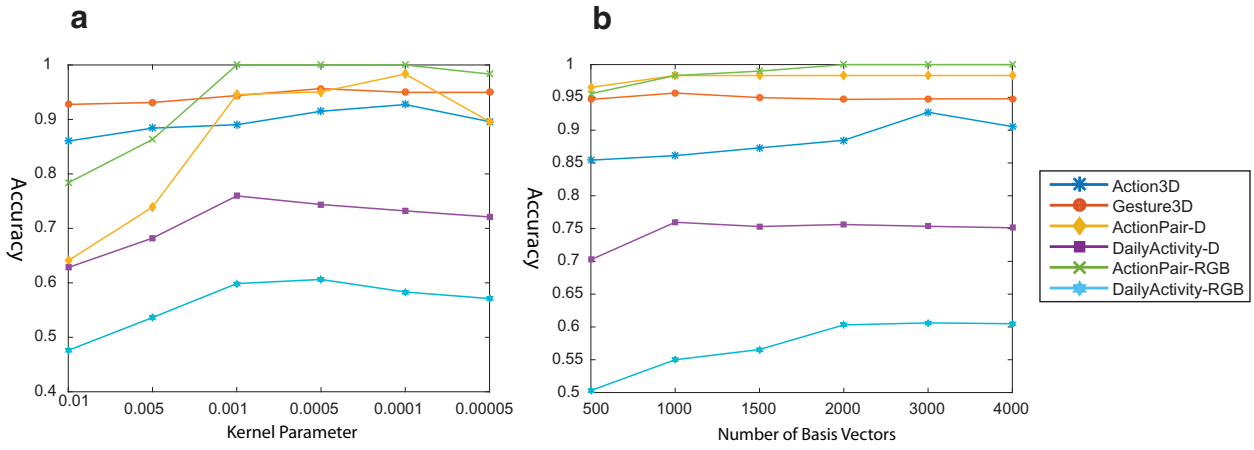

Fig. 4. Accuracy of classification of various datasets with changing CKSVD (a) kernel parameter and (b) number of basis vectors. "-D" means the depth data in the dataset and "-RGB" means the RGB data.

Table 1
Recognition accuracy (%) of different methods on MSR action 3D dataset.

Method	Accuracy (%)
Dynamic temporal warping [36]	54.00
Hidden Markov model [37]	63.00
Action graph on bag of 3D points [8]	74.40
Motion maps-based HOG [38]	85.52
Mining actionlet ensemble [4]	88.20
HON4D [27]	88.36
Our method (BOW)	**78.35**
Our method (EMK)	**92.73**
Our method (HKD)	**93.99**

Table 2
Recognition accuracy (%) of different methods on MSR gesture 3D dataset.

Method	Accuracy (%)
SVM on raw features	62.77
High dimensional convolutional network [39]	69
Action graph on occupancy features [40]	80
Action graph on silhouette features [40]	87.7
Random occupancy patterns [6]	88.5
HON4D [27]	92.45
Our method (BOW)	**84.12**
Our method (EMK)	**95.66**
Our method (HKD)	**96.09**

test data. Fig. 6 presents the confusion matrix. Confusions mainly occur between "hammer" and "forward punch", "hand catch" and "high arm wave", and "tennis serve" and "pick up and throw" due to similar motions.

Table 1 shows the accuracy comparison of different methods. Our approaches outperforms state-of-the-art methods. Note that the results from [27] is not as the same as published paper because their experiment is not performed with the exact same setup as other publications. We acquire their code and run the test with standard setup on dividing test and train data which leads to accuracy of 88.36%.

Table 1 indicates that our approach using HKD outperforms the one with EMK as the high-layer kernel descriptor. The underlying reason is that HKD captures spatio-temporal structure information while EMK does not. With structure information, HKD is able to differentiate motion of different body parts and give different importance to them in feature summarization. This will generate more discriminative feature descriptors, and thereby improving the recognition performance.

To show our method is general in terms of choosing test and train data and is not over fitted on standard experiment setup, we also run it on all possible permutations of having half of subjects for train data and the other half for test data. Result is 252 runs for choosing five subjects out of ten. We run the experiments with the same set of hyperparameters, and obtain an accuracy of $82.40 \pm 3.63\%$ (mean $\pm$ std). This shows that our method does not depend on a specific permutation for choosing train and test data.

4.4. Results on MSR gesture 3D dataset

The confusion matrix corresponding to our best result on this dataset is shown in Fig. 7. The best result is 96.09% which is achieved by our method where we employ HKD in hierarchical structure. Confusions mainly occur between "past" and "hungry", "finish" and "milk", "j" and "blue", and "green" and "store". Examples of misclassification are illustrated in Fig. 5.

Our method is also compared with existing methods [6,27,39,40]. Table 2 addresses our performance on this dataset compared to the state-of-the-art methods. Results indicate that our method significantly outperforms existing methods. The proposed 3D gradient kernel can better summarize local motion information in 3D patch. In addition, both EMK and HKD are capable of learning nonlinear correlations of human body parts, and provide rich mid-level features for classification. However, HKD still outperforms EMK due to the capability of capturing spatio-temporal structure information.

4.5. Results on MSR action pairs dataset

Table 3 shows the accuracy comparison between our method and previous work [4,27,38] as reported in [27]. We also achieve state-of-the-art performance on MSR Action Pairs dataset. As we have achieved superior performance of 100% on this dataset, we do not present the confusion matrix. Capturing the temporal changes in our 3D kernel descriptor gives this method the ability to differentiate between 3D patches with the same shape and different motion direction.

We also test our approach given one modality (RGB or depth) data, and show the recognition accuracy in Table 3. Results show that our approach is very effective even one modality is given. Using EMK kernel descriptor, our approach achieves 98.33% accuracy on depth-only data, and achieves 100% accuracy when RGB-only data and RGB-D data are given. This also demonstrates the benefit of using both modalities for recognition. Using HKD, our approach achieves 100% accuracy on depth-only, RGB-only, and RGB-D data since HKD captures spatio-temporal structure information compared with EMK.

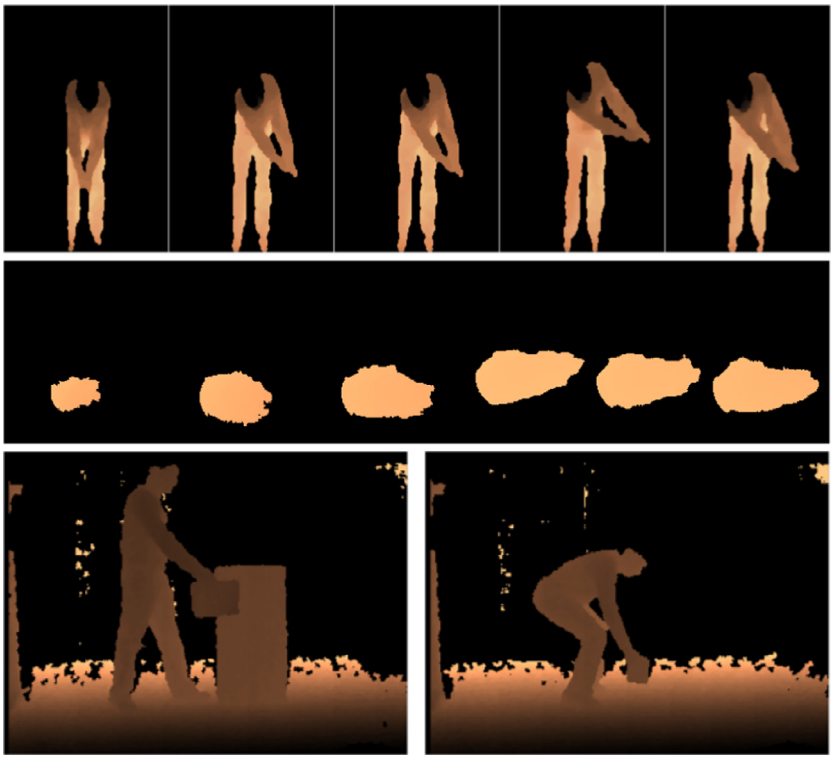

Fig. 5. Examples of misclassified instances.

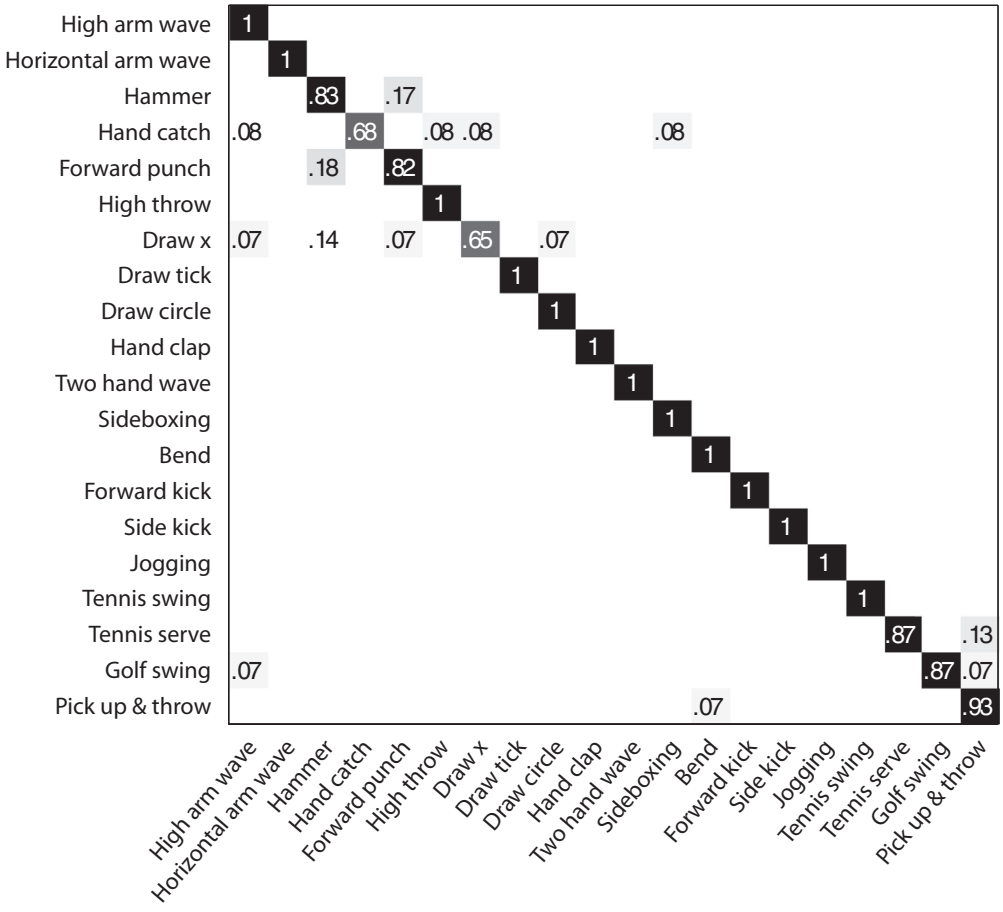

Fig. 6. Confusion matrix for best result on MSR Action 3D dataset.

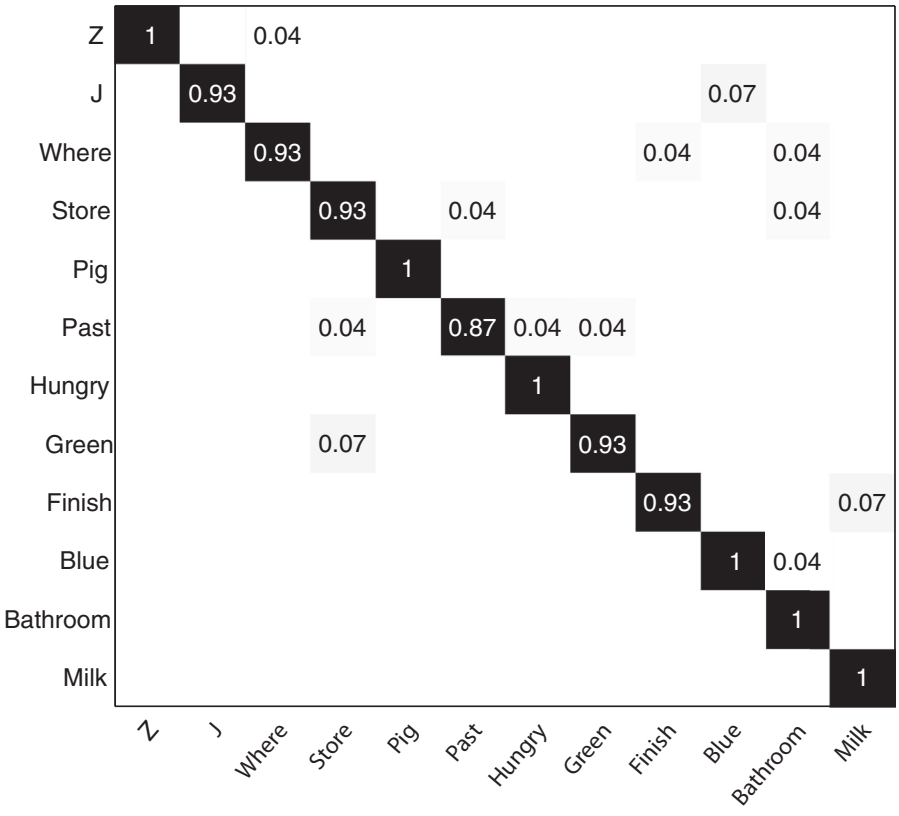

Fig. 7. Confusion matrix of our method using HKD as high-level kernel descriptor on MSR Gesture 3D dataset. Overall accuracy is 96.09%.

Table 3
Recognition accuracy (%) of different methods on MSR action pairs dataset.

Method	Accuracy (%)
Motion maps-based HOG [38]	66.11
Mining actionlet ensemble [4]	82.22
HON4D [27]	96.67
Our method (BOW - Depth)	**81.90**
Our method (BOW - RGB)	**77.67**
Our method (BOW - RGB-D)	**85.19**
Ourmethod (EMK - Depth)	**98.33**
Our method (EMK - RGB)	**100**
Our method (EMK - RGB-D)	**100**
Our method (HKD - Depth)	**100**
Our method (HKD - RGB)	**100**
Our method (HKD - RGB-D)	**100**

Table 4
Recognition accuracy (%) of different methods on MSR Daily Activity 3D dataset. Note that the details about training/testing split in these methods are not reported, and thus the recognition accuracy cannot be directly compared. In this paper, we report recognition accuracy using one person out cross validation scheme.

Method	Accuracy (%)
Deep motion maps [38]	43.13
RGGP [26]	72.1
Moving pose [41]	73.8
Local HON4D [27]	80.0
Actionlet ensemble [4]	85.75
Our method (BOW - Depth)	**61.75**
Our method (BOW - RGB)	**47.11**
Our method (BOW - RGB-D)	**62.53**
Our method (EMK - Depth)	**75.93**
Our method (EMK - RGB)	**60.62**
Our method (EMK - RGB-D)	**83.13**
Our method (HKD - Depth)	**68.21**
Our method (HKD - RGB)	**50.31**
Our method (HKD - RGB-D)	**73.21**

4.6. Results on MSR daily activity dataset

This dataset is designed to represent human daily activities in the living room. Having two different poses with presenting human-object interaction make it more challenging than other datasets and close to real world examples. The best performance result on this dataset belongs to the part where we use EMK on both RGB and depth sequences. The confusion matrix corresponding to our best result on this dataset is shown in Fig. 8, where the overall accuracy is 83.13%.

We also list recognition results of previous methods in Table 4. Note that all these methods do not report details about training/testing data split. Therefore, it is not possible for us to conduct a fair comparison between our results and all these methods. In this paper, we report recognition accuracy using one person out cross validation scheme.

Recognition accuracy of our method given RGB-only, depth-only data, and RGB-D data are also reported in Table 4. Results demonstrate the effectiveness of using both RGB and depth in recognition as using RGB-D data significantly boosts the recognition performance using either EMK or HKD descriptor. The recognition performance between EMK and HKD is also compared here. It is interesting to see that HKD performs worse than EMK, which is inconsistent with results in other three datasets used in this paper. The underlying reason is that the MSR Daily Activity dataset has extremely noisy background, which could degrade the performance if the structure information of noisy background is also considered. The reason that we have shown both EMK and HKD methods here is to show that in some cases EMK still has the advantage. Our main contribution is using the low-level

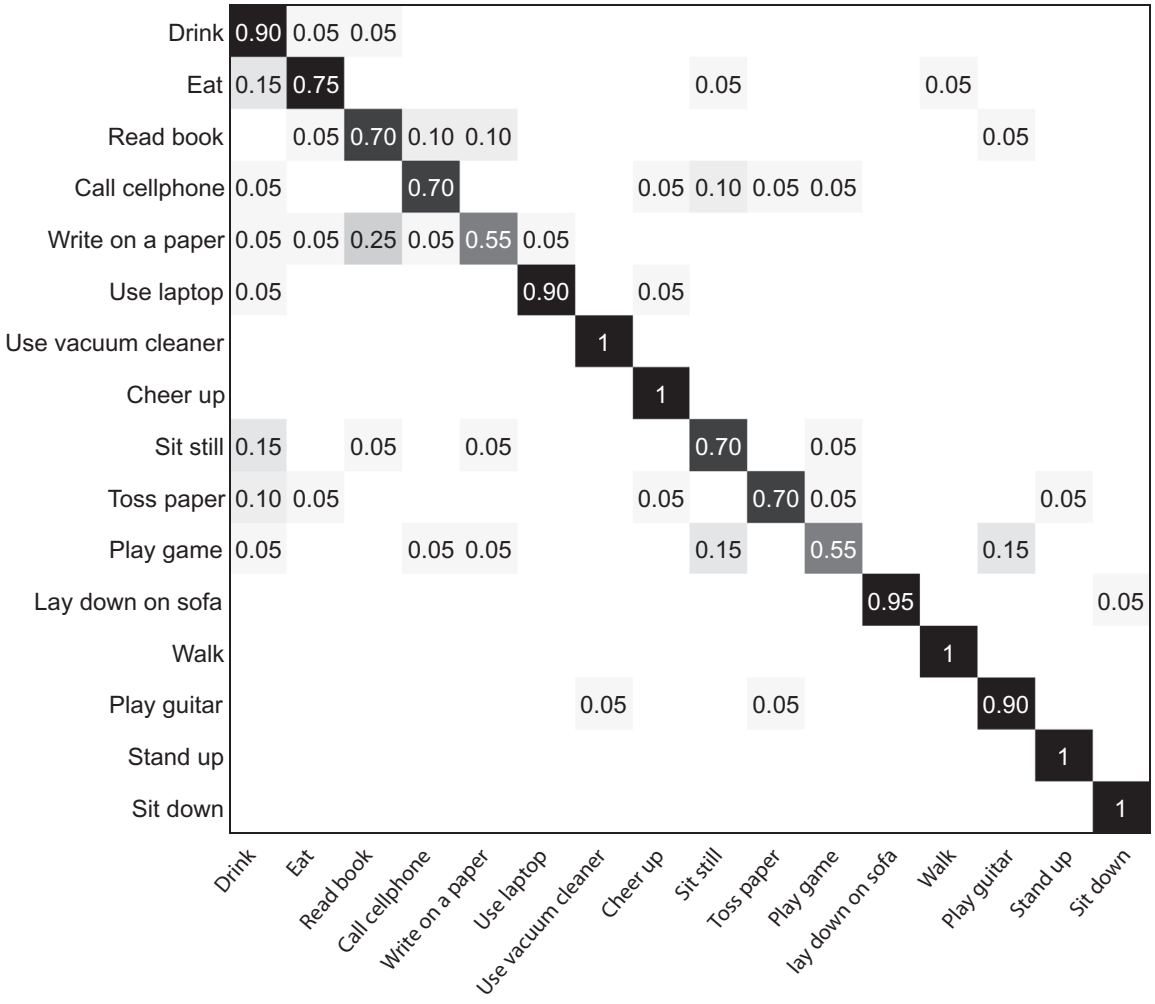

Fig. 8. Confusion matrix of our approach on MSR Daily Activity 3D dataset using EMK descriptor. Overall accuracy is 83.13%.

features by employing kernel descriptor and for higher level of our framework you still have the choice of various methods and it really depends on type of data which you are working on.

5. Conclusion

We have proposed a simple and effective method for employing kernel descriptors in describing the 3D geometric surface of depth sequences (oriented normal vectors) [7] in human action recognition task. Our descriptor is an extension of gradient kernel descriptor in [2] which is shown to be an effective way to capture pixel-level attributes in object recognition. In next level we use EMK and HKD to abstract mid-level features to produce video-level representations. Through extensive experiments we show how our method outperform previous methods and achieve state-of-the-art performance in standard action recognition RGB-D sequences.

Acknowledgments

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, NPS award N00244-15-1-0041, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

[1] Y. Kong, B. Satarboroujeni, Y. Fu, Hierarchical 3D Kernel descriptors for action recognition using depth sequences, in: Proceedings of the FG, 2015.
[2] L. Bo, X. Ren, D. Fox, Kernel descriptors for visual recognition, in: Proceedings of the NIPS, 2010.
[3] L. Bo, C. Sminchisescu, Efficient match Kernel between sets of features for visual recognition, in: Proceedings of the NIPS, 2009.
[4] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on CVPR, 2012, 2012, pp. 1290–1297.
[5] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Proceedings of the CVPR workshop, 2010.
[6] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3d action recognition with random occupancy patterns, in: Proceedings of the 12th ECCV - Volume Part II, ECCV'12, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 872–885.
[7] S. Tang, X. Wang, X. Lv, T.X. Han, J. Keller, Z. He, M. Skubic, S. Lao, Histogram of oriented normal vectors for object recognition with a depth sensor, in: Proceedings of the 11th ACCV - Volume Part II, ACCV'12, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 525–538.
[8] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Proceedings of the IEEE Computer Society Conference on CVPRW, 2010, 2010, pp. 9–14.
[9] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the VS-PETS, 2005, pp. 65–72.
[10] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: Proceedings of the ICPR, 2004, pp. 32–36.
[11] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2) (2005) 107–123.
[12] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of the BMVC, 2008, pp. 1–10.
[13] W. Li, N. Vasconcelos, Recognizing activities by attribute dynamics, in: Proceedings of the NIPS, 2012, pp. 1106–1114.
[14] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: Proceedings of the CVPR, 2008, pp. 1–8.
[15] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: Proceedings of the ICCV, 2009, pp. 1593–1600.
[16] M. Raptis, L. Sigal, Poselet key-framing: a model for human activity recognition, in: Proceedings of the CVPR, IEEE, 2013, pp. 2650–2657.

[17] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[18] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: Proceedings of the CVPR, 2011, pp. 3337–3344.
[19] Y. Wang, G. Mori, Hidden part models for human action recognition: probabilistic vs. max-margin, IEEE TPAMI 33 (7) (2011) 1310–1323.
[20] Y. Kong, Y. Jia, Y. Fu, Interactive phrases: Semantic descriptions for human interaction recognition, in: Proceedings of the IEEE TPAMI, 2014.
[21] W. Choi, K. Shahid, S. Savarese, Learning context for collective activity recognition, in: Proceedings of the CVPR, 2011, pp. 3273–3280.
[22] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3D action recognition with random occupancy patterns, in: Proceedings of the ECCV, 2012a, pp. 872–885.
[23] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the CVPR, 2012b, pp. 1290–1297.
[24] S. Hadfield, R. Bowden, Hollywood 3D: recognizing actions in 3D natural scenes, in: Proceedings of the CVPR, 2013, pp. 3398–3405.
[25] L. Xia, J.K. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Proceedings of the CVPR, 2013, pp. 2834–2841.
[26] L. Liu, L. Shao, Learning discriminative representations from rgb-d video data, in: Proceedings of the IJCAI, 2013, pp. 1493–1500.
[27] O. Oreifej, Z. Liu, HON4D: histogram of oriented 4D normals for activity recognition from depth sequences, in: Proceedings of the CVPR, 2013, pp. 716–723.
[28] C. Jia, Y. Kong, Z. Ding, Y. Fu, Latent tensor transfer learning for rgb-d action recognition, in: Proceedings of the ACM Multimedia, 2014.
[29] L. Chen, W. Li, D. Xu, Recognizing rgb images by learning from rgb-d data, in: Proceedings of the CVPR, 2014.
[30] X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences, in: Proceedings of the CVPR, 2014.
[31] B. Schölkopf, A. Smola, E. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[32] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[33] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on CVPR 2005, 1, 2005, pp. 886–893.
[34] D. Haussler, Convolution kernels on discrete structures, 1999.
[35] I. Laptev, T. Lindeberg, Space-time interest points, in: Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003, 1, 2003, pp. 432–439.
[36] M. Müller, T. Röder, Motion templates for automatic classification and retrieval of motion capture data, in: Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Aanimation, Eurographics Association, 2006, pp. 137–146.
[37] F. Lv, R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, in: Proceedings of the ECCV, 2006.
[38] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the 20th ACM MM, MM '12, ACM, New York, NY, USA, 2012, pp. 1057–1060.
[39] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell., 35 (1) (2013) 221–231.
[40] A. Kurakin, Z. Zhang, Z. Liu, A real time system for dynamic hand gesture recognition with a depth sensor, in: Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, 2012, pp. 1975–1979.
[41] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection, in: Proceedings of the ICCV, 2013.
[42] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: Proceedings of the CVPR, 2010.
[43] A.A. Chaaraouia, J.R. Padilla-Lópeza, P. Climent-Pérezb, F. Flórez-Revueltab, Evolutionary joint selection to improve human action recognition with RGB-D devices, Expert Syst. Appl. 41 (3) (2014) 786–794.
[44] Y. Zhu, W. Chen, G. Guo, Fusing spatiotemporal features and joints for 3d action recognition, in: Proceedings of the CVPR Workshop, 2013.
[45] E. Ohn-Bar, M.M. Trivedi, Joint angles similarities and HOG2 for action recognition, in: Proceedings of the CVPR Workshop, 2013.